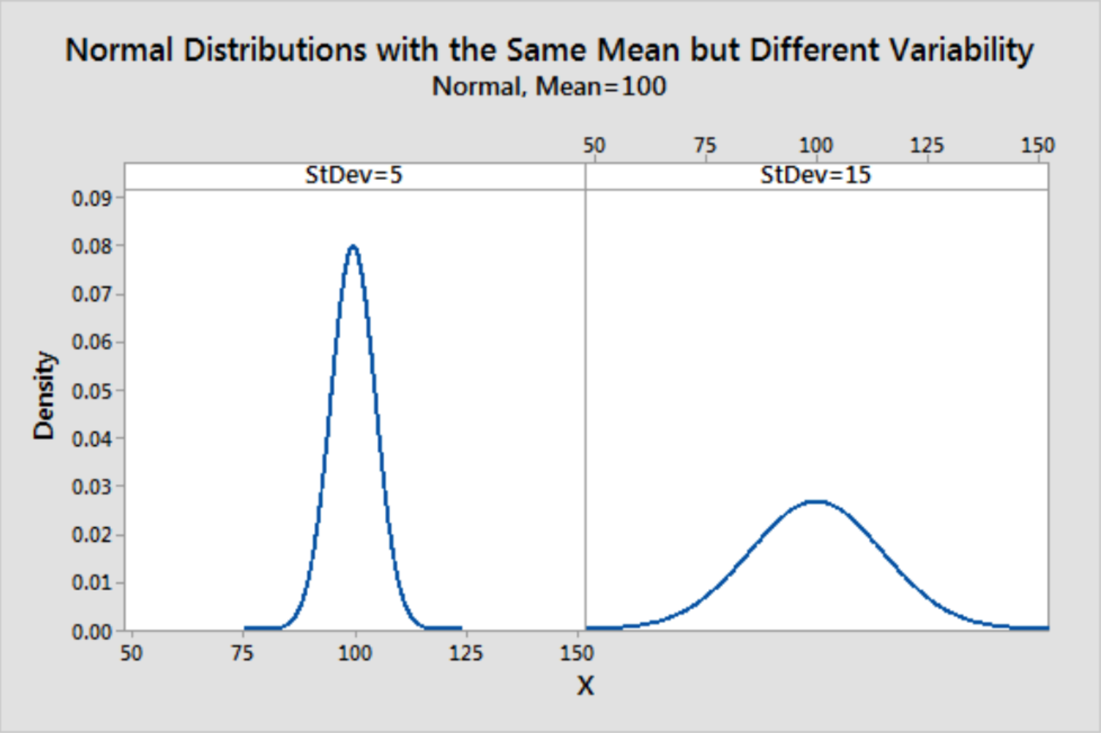
Measures of Variability: Range, Interquartile Range, Variance, and Standard Deviation

By Jim Frost — 75 Comments

A measure of variability is a summary statistic that represents the amount of dispersion in a dataset. How spread out are the values? While a measure of central tendency describes the typical value, measures of variability define how far away the data points tend to fall from the center. We talk about variability in the context of a distribution of values. A low dispersion indicates that the data points tend to be clustered tightly around the center. High dispersion signifies that they tend to fall further away.

In statistics, variability, dispersion, and spread are synonyms that denote the width of the distribution. Just as there are multiple measures of central tendency, there are several measures of variability. In this blog post, you'll learn why understanding the variability of your data is critical. Then, I explore the most common measures of variability—the range, interquartile range, variance, and standard deviation. I'll help you determine which one is best for your data.

The two plots below show the difference graphically for distributions with the same mean but more and less dispersion. The panel on the left shows a distribution that is tightly clustered around the average, while the distribution in the right panel is more spread out.



Why Understanding Variability is Important

Let's take a step back and first get a handle on why understanding variability is so essential. Analysts frequently use the mean to summarize the center of a population or a process. While the mean is relevant, people often react to variability even more. When a distribution has lower variability, the values in a dataset are more consistent. However, when the variability is higher, the data points are more dissimilar and extreme values become more likely. Consequently, understanding variability helps you grasp the likelihood of unusual events.

In some situations, extreme values can cause problems! Have you seen a weather report where the meteorologist shows extreme heat and drought in one area and flooding in another? It would be nice to average those together! Frequently, we feel discomfort at the extremes more than the mean. Understanding that variability around the mean provides critical information.

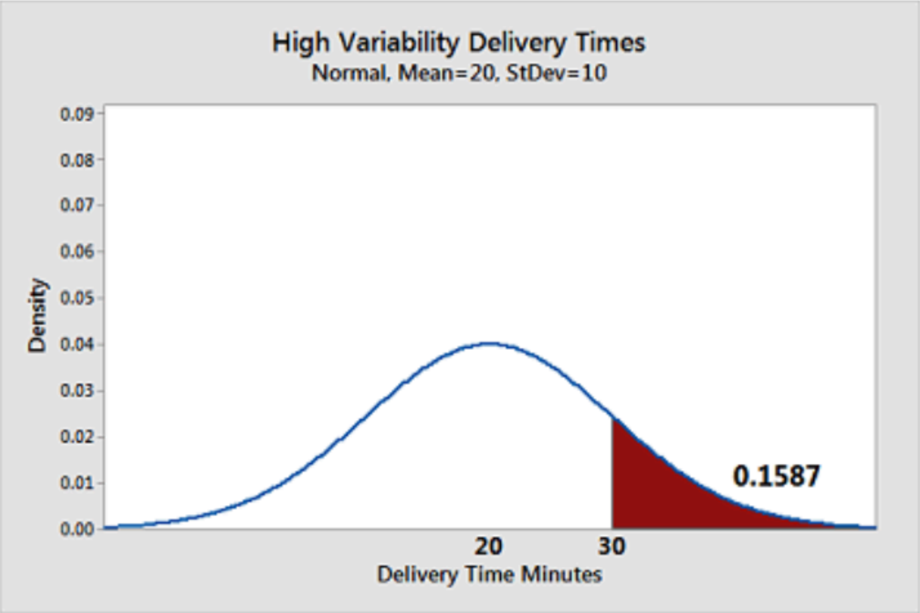
Variability is everywhere. Your commute time to work varies a bit every day. When you order a favorite dish at a restaurant repeatedly, it isn't exactly the same each time. The parts that come off an assembly line might appear to be identical, but they have subtly different lengths and widths.

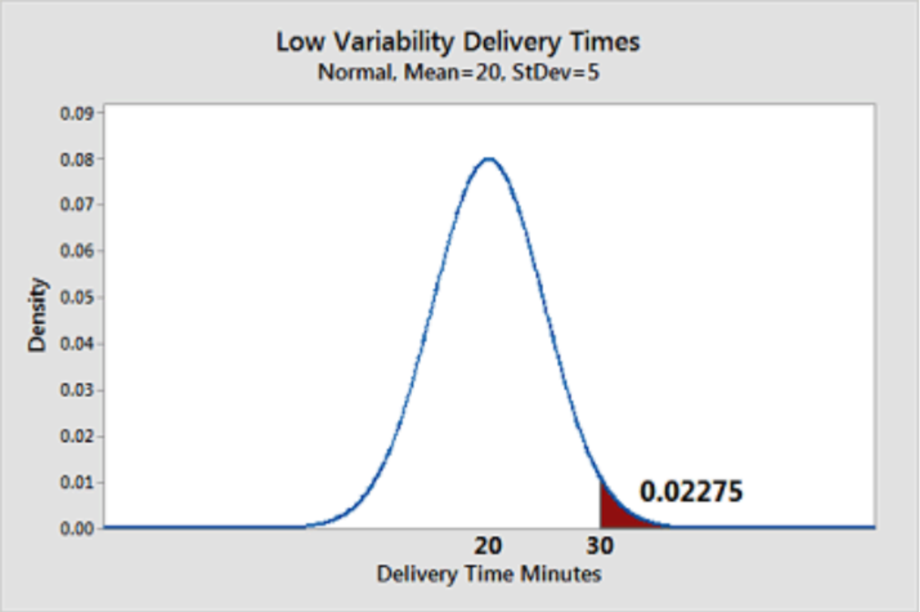
These are all examples of real-life variability. Some degree of variation is unavoidable. However, too much inconsistency can cause problems. If your morning commute takes much longer than the mean travel time, you will be late for work. If the restaurant dish is much different than how it is usually, you might not like it at all. And, if a manufactured part is too much out of spec, it won't function as intended.

Some variation is inevitable, but problems occur at the extremes. Distributions with greater variability produce observations with unusually large and small values more frequently than distributions with less variability.

Suppose we study their delivery times, calculate the variability for each place, and determine that their variabilities are different. We've computed the standard deviations for both restaurants—which is a measure that we'll come back to later in this post. How significant is this difference in getting pizza to their customers promptly?

The graphs below display the distribution of delivery times and provide the answer. The restaurant with more variable delivery times has the broader distribution curve. I've used the same scales in both graphs so you can visually compare the two distributions.



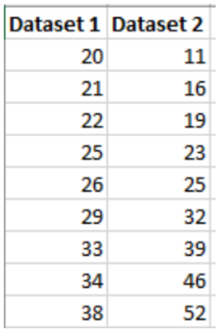


In these graphs, we consider a 30-minute wait or longer to be unacceptable. We're hungry after all! The shaded area in each chart represents the proportion of delivery times that surpass 30 minutes. Nearly 16% of the deliveries for the high variability restaurant exceed 30 minutes. On the other hand, only 2% of the deliveries take too long with the low variability restaurant. They both have an average delivery time of 20 minutes, but I know where I'd place my order when I'm hungry!

As this example shows, the central tendency doesn't provide complete information. We also need to understand the variability around the middle of the distribution to get the full picture. Now, let's move on to the different ways of measuring variability!

Range

Let’s start with the range because it is the most straightforward measure of variability to calculate and the simplest to understand. The range of a dataset is the difference between the largest and smallest values in that dataset. For example, in the two datasets below, dataset 1 has a range of 20 – 38 = 18 while dataset 2 has a range of 11 – 52 = 41. Dataset 2 has a broader range and, hence, more variability than dataset 1.



Dataset 1	Dataset 2
20	11
21	16
22	19
25	23
26	25
29	32
33	39
34	46
38	52

While the range is easy to understand, it is based on only the two most extreme values in the dataset, which makes it very susceptible to outliers. If one of those numbers is unusually high or low, it affects the entire range even if it is atypical.

Additionally, the size of the dataset affects the range. In general, you are less likely to observe extreme values. However, as you increase the sample size, you have more opportunities to obtain these extreme values. Consequently, when you draw random samples from the same population, the range tends to increase as the sample size increases. Consequently, use the range to compare variability only when the sample sizes are similar.

Variance

Variance is the average squared difference of the values from the mean. Unlike the previous measures of variability, the variance includes all values in the calculation by comparing each value to the mean. To calculate this statistic, you calculate a set of squared differences between the data points and the mean, sum them, and then divide by the number of observations. Hence, it’s the average squared difference.

There are two formulas for the variance depending on whether you are calculating the variance for an entire population or using a sample to estimate the population variance. The equations are below, and then I work through an example in a table to help bring it to life.

Population variance

The formula for the variance of an entire population is the following:

$$\sigma^2 = \frac{\sum (X - \mu)^2}{N}$$

In the equation, σ^2 is the population parameter for the variance, μ is the parameter for the population mean, and N is the number of data points, which should include the entire population.

Sample variance

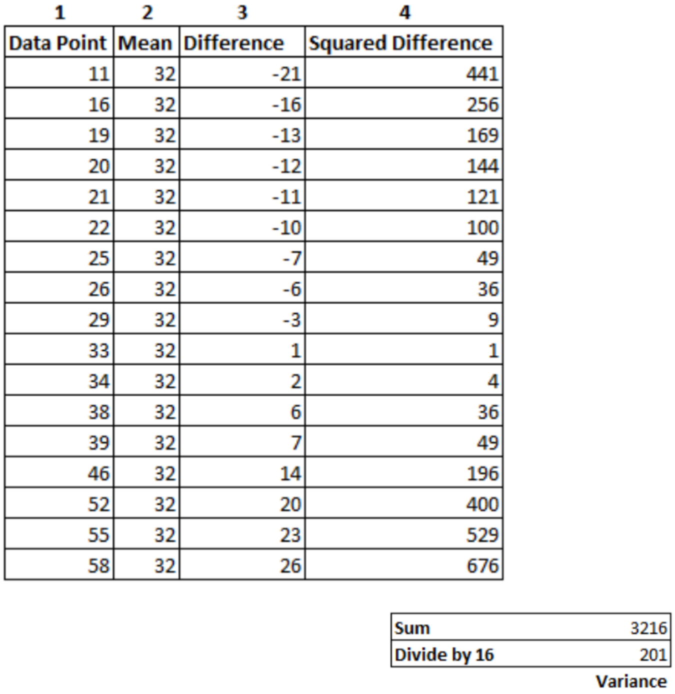
To use a sample to estimate the variance for a population, use the following formula. Using the previous equation with sample data tends to underestimate the variability. Because it's usually impossible to measure an entire population, statisticians use the equation for sample variances much more frequently.

$$s^2 = \frac{\sum (X - M)^2}{N - 1}$$

In the equation, s^2 is the sample variance, and M is the sample mean. N-1 in the denominator corrects for the tendency of a sample to underestimate the population variance.

Example of calculating the sample variance

I'll work through an example using the formula for a sample on a dataset with 17 observations in the table below. The numbers in parentheses represent the corresponding table column number. The procedure involves taking each observation (1), subtracting the sample mean (2) to calculate the difference (3), and squaring that difference (4). Then, I sum the squared differences at the bottom of the table. Finally, I take the sum and divide by 16 because I'm using the sample variance equation with 17 observations (17 – 1 = 16). The variance for this dataset is 201.



1	2	3	4
Data Point	**Mean**	**Difference**	**Squared Difference**
11	32	-21	441
16	32	-16	256
19	32	-13	169
20	32	-12	144
21	32	-11	121
22	32	-10	100
25	32	-7	49
26	32	-6	36
29	32	-3	9
33	32	1	1
34	32	2	4
38	32	6	36
39	32	7	49
46	32	14	196
52	32	20	400
55	32	23	529
58	32	26	676

Sum	3216
Divide by 16	201
	Variance

Because the calculations use the squared differences, the variance is in squared units rather the original units of the data. While higher values of the variance indicate greater variability, there is no intuitive interpretation for specific values. Despite this limitation, various statistical tests use the variance in their calculations. For an example, read my post about the F-test and ANOVA.

While it is difficult to interpret the variance itself, the standard deviation resolves this problem!

Standard Deviation

The standard deviation is the standard or typical difference between each data point and the mean. When the values in a dataset are grouped closer together, you have a smaller standard deviation. On the other hand, when the values are spread out more, the standard deviation is larger because the standard distance is greater.

Conveniently, the standard deviation uses the original units of the data, which makes interpretation easier. Consequently, the standard deviation is the most widely used measure of variability. For example, in the pizza delivery example, a standard deviation of 5 indicates that the typical delivery time is plus or minus 5 minutes from the mean. It’s often reported along with the mean: 20 minutes (s.d. 5).

The standard deviation is just the square root of the variance. Recall that the variance is in squared units. Hence, the square root returns the value to the natural units. The symbol for the standard deviation as a population parameter is σ while s represents it as a sample estimate. To calculate the standard deviation, calculate the variance as shown above, and then take the square root of it. Voila! You have the standard deviation!

In the variance section, we calculated a variance of 201 in the table.

$$\sqrt{201} = 14.177$$

Therefore, the standard deviation for that dataset is 14.177.

The Empirical Rule for the Standard Deviation of a Normal Distribution

When you have normally distributed data, or approximately so, the standard deviation becomes particularly valuable. You can use it to determine the proportion of the values that fall within a specified number of standard deviations from the mean. For example, in a normal distribution, 68% of the values will fall within +/- 1 standard deviation from the mean. This property is part of the Empirical Rule. This rule describes the percentage of the data that fall within specific numbers of standard deviations from the mean for bell-shaped curves.

Mean +/- standard deviations	Percentage of data contained
1	68%
2	95%
3	99.7%

Let's take another look at the pizza delivery example where we have a mean delivery time of 20 minutes and a standard deviation of 5 minutes. Using the Empirical Rule, we can use the mean and standard deviation to determine that 68% of the delivery times will fall between 15-25 minutes (20 +/- 5) and 95% will fall between 10-30 minutes (20 +/- 2*5).

Which is Best—the Range, Interquartile Range, or Standard Deviation?

First off, you probably notice that I didn't include the variance as one of the options in the heading above. That's because the variance is in squared units and doesn't provide an intuitive interpretation. So, I've crossed that off the list. Let's go over the other three measures of variability.

When you are comparing samples that are the same size, consider using the range as the measure of variability. It's a reasonably intuitive statistic. Just be aware that a single outlier can throw the range off. The range is particularly suitable for small samples when you don't have enough data to calculate the other measures reliably, and the likelihood of obtaining an outlier is also lower.

When you have a skewed distribution, the median is a better measure of central tendency, and it makes sense to pair it with either the interquartile range or other percentile-based ranges because all of these statistics divide the dataset into groups with specific proportions.

For normally distributed data, or even data that aren't terribly skewed, using the tried and true combination reporting the mean and the standard deviation is the way to go. This combination is by far the most common. You can still supplement this approach with percentile-base ranges as you need.